通过提示工程提升人工智能：来自数据科学应用的见解与案例研究

Snehasish Paul[1*]、Rohit Kumar[2] 和 Laxman Das[3]

[1,2] 德里大学图书馆与信息科学系，印度德里，邮编：110007 ，
印度[3]共生商学院诺伊达校区，共生国际大学（视同大学）浦那校区，印度，邮编：201301

***通讯作者**－snehasishpaulas98@gmail.com

ORCID ID- https://orcid.org/0009-0003-2730-5314

摘要

提示工程正逐渐成为人工智能领域不可或缺的重要组成部分，并改变着数据科学家在分析应用中与大语言模型（LLM）的交互方式。本文分享了关于提示工程的方法论、有效性及应用的多项实证研究结果。通过医疗保健、材料科学、金融服务和商业智能领域的案例研究，我们展示了结构化提示技术如何提升各类任务的表现，提升幅度从 6% 到 30% 以上不等。研究结果表明，提示词的有效性取决于其复杂程度。此外，模型架构和优化策略也受这些因素影响。我们还发现，思维链推理和自动优化器等先进框架前景可观。证据表明，提示工程有助于实现强大的人工智能本地化。不过，在标准化、可解释性以及人工智能的伦理使用方面，仍有大量问题有待探讨。

关键词：人工智能，数据科学、提示工程、机器学习、大语言模型（LLM）

引言

大语言模型正在改变计算机理解和使用人类语言进行交流的方式。过去，人们通过修改架构、大规模微调以及针对特定领域进行训练来提升人工智能的性能。由于这些过程十分复杂，实际操作需要大量计算资源和专业知识。通用语言模型在预训练过程中获得了丰富的知识和推理机制。提示工程设计能够揭示嵌入这些神经网络模式中的知识和推理机制。近年来的技术不再将提示词视为简单的问题，而是将其视为一种接口，模型可以通过它专注于一组特定知识，同时限制输出空间，以满足预期任务的要求。基于这一认识，研究人员提出了多种方法，包括零样本与少样本学习、思维链、检索增强学习和自动化优化框架。在某个领域，数据的使用为通过提示工程进行创新提供了大量机会（Ge等人，2025）。这是因为该领域涉及许多分析任务。此外，这些任务通常具有领域专属性，且要求复杂。传统的限制还使非技术利益相关者难以利用人工智能。多年来，将原始数据转化为有意义的数据往往需要高水平的编程技能和专业领域知识。借助提示工程，人们可以用自然语言说明分析目标，从而让更多人能够在数据科学工作流程中发挥实际作用。目前，研究人员正在开展一项研究，考察各种提示工程技术，以及它们分别适用于哪些数据科学场景。同样，研究也将探讨哪种类型的提示词最适合哪种类型的分析。最后，研究将分析最速下降法或其他形式的优化框架，探究哪种方法能在性能与复杂度之间实现最佳权衡。

提示工程的方法论基础

如今，提示工程涵盖了许多技术，这些技术在复杂程度、计算需求和应用场景方面各不相同。推理提示是提示工程方法的一项有益补充。思维链提示（CoTP）是指在现有模型的基础上加入推理结构，使其能够进行多步问题求解。CoTP意味着逻辑路径会不断分支，从而能够在沿着初始路径求解的同时探索其他解题策略。通常，对于需要较高层次推理或数学能力的任务，采用能够支持结构化推理的框架会有所帮助。根据Khatuya等人（2025）的研究，他们提出的FINDER框架结合FINDER检索与程序化提示后，在FinQA和ConvFinQA基准数据集上的执行准确率分别提高了5.98和4.05个百分点。这表明，将推理结构直接融入模型设计，可以提升模型在需要逻辑序列的任务上的表现。

指令提示会详细说明具体任务及输出格式。这类提示采用直接的语言表达期望的行为，并限制输出规范。Hu 等人（2024）通过开发一个由基线任务描述、依据指南制定的指令、用于错误分析的指令以及经过标注的少样本示例组成的四组件框架，提供了结构化指令提示在临床命名实体识别任务中具有实用价值的证据。研究结果表明，系统地应用所有提示组件后，GPT-4宽松F1分数在MTSamples数据集上从0.804升至0.861，在VAERS数据集上则从0.593升至0.736。这些显著提升表明，良好的指令设计应在适当时机考虑多个方面的规范要求。

上下文提示强调，除了任务本身的信息，还应向模型提供更广泛的上下文、领域知识和示例。在某些学科中，正确答案取决于专业词汇、文化或历史背景，这种方法很有用。根据Polak & Morgan（2024）的研究，ChatExtract系统采用对话提示来提取材料数据。先进对话模型对材料属性的准确率和召回率指标接近90%。正如该方法所示，使用丰富且可通过迭代验证的上下文信息，可以改进工程领域中的信息提取。

使用先进的自动化优化系统，可以系统地通过 Google 搜索探索各种提示设计空间，以改善循环神经网络的表现。Lieander 等人（2025）提出了 PO2G，这是一种基于梯度的优化机制，全称为“双梯度提示优化”（Prompt Optimization with Two Gradients）。经过总计三次迭代，该方法的准确率约为 89%。这一表现优于 ProTeGi 基线，后者需要六次迭代才能达到相同水平。我们可以优化算法，让提示词的处理速度快于人类所能达到的速度。根据 Agarwal 等人（2024）的研究，PromptWizard 是一种智能体驱动框架，可对指令和示例进行迭代变异，并结合基于批评器的评估。在 35 项以上的评估任务中，PromptWizard 始终优于现有的提示策略。

不同人工智能模型和任务的比较效果

提示工程技术对某些模型架构和任务可能效果显著，但对其他架构和任务则未必如此。对于希望为特定用途选择模型和提示策略的从业者而言，了解性能差异十分重要。近期研究通过在越来越多的模型和基准上开展实证评估，帮助我们深入了解这些关系。

表1：真实世界提示工程研究中的关键模型性能指标 h

模型	任务领域	性能指标	来源
GPT-4	临床命名实体识别	宽松F1：0.861（MTSamples），0.736（VAERS）	（Hu等，2024年）
GPT-3.5-turbo	职位分类	+6% 95%召回率下的精确率与监督式方法	（Clavié等人，2023年）
GPT-3.5-turbo / Claude 2	网络钓鱼检测	F1值：92.74%（提示方法）	（Trad 和 Chehab，2024年）
GPT-4	Data 预处理	在 12 个数据集中有 4 个达到 100% 的准确率/F1值	（Zhang 等人，2023年）
GPT-4	模式匹配	召回率：100%（DeepMDatasets），91.8%（Fabricated）	（Feng 等，2024年）
Claude	结构化数据任务	任务集中的最高准确率	（Schmidt等人，2025）
ChatGPT-4o	结构化生成	最节省词元和时间	(Schmidt等人，2025)

表1展示了实证研究中精选的性能亮点，这些研究考察了提示词在不同模型和任务中的有效性。根据数据，关于模型能力和提示响应性呈现出若干趋势。

Hu 等人（2024年）在临床命名实体识别方面的结果表明，GPT-4 等神经网络对结构化提示组件的响应比其前代模型更有效。应用系统提示框架后，模型的 F1值显著提升，这表明复杂模型具备可通过提示机制激发的潜在能力。然而，基线与优化提示之间仍存在较大的性能差距，这说明要发挥这些能力，精心设计提示至关重要，而不能仅仅依赖复杂模型本身。

正如 Clavié 等人（2023）所示，精心设计的零样本提示词能够让 GPT-3.5-turbo 在职位分类任务中胜过监督学习方法。事实上，与表现最佳的监督基线相比，其95%召回率下的精确率提高了6%。这些发现挑战了人们普遍认为模型要取得成功就需要任务特定训练数据的看法。我们发现，在某些情况下，提示工程可以作为监督学习流水线的可行替代方案。这具有重要的实际意义，因为组织或许无需投入资源收集标注训练数据和进行模型微调，就能部署有效的分类系统。

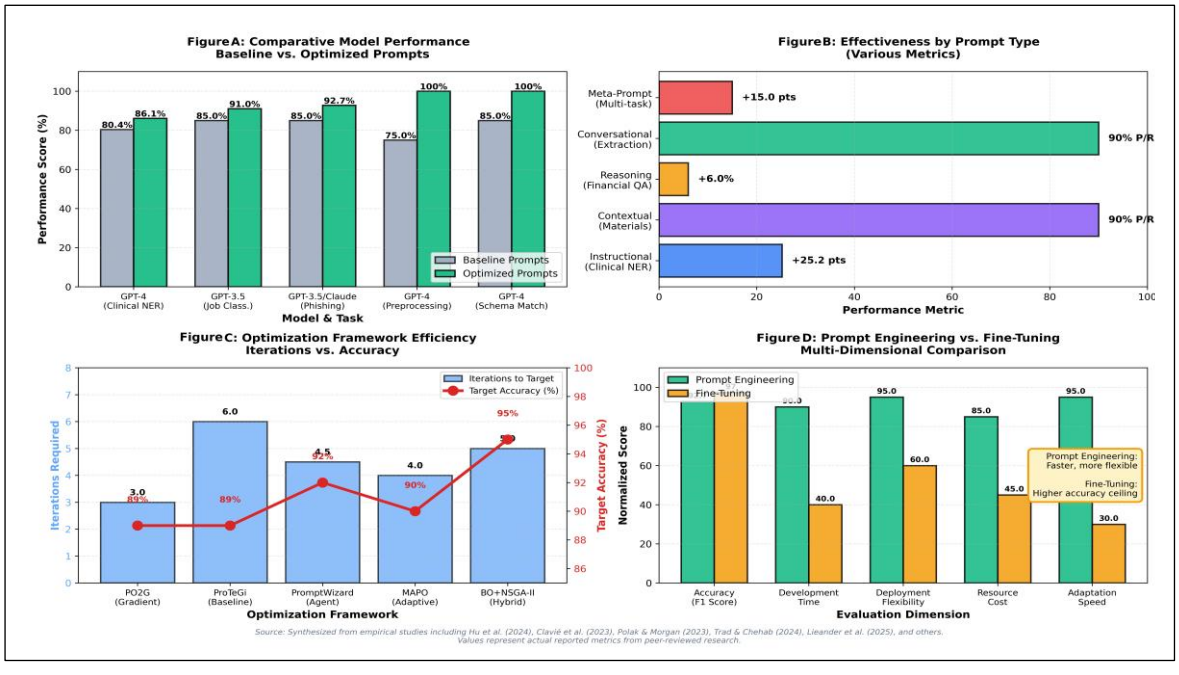


图 *1*：数据科学应用中提示工程的全面效果分析

图1的A面板展示了五种不同模型-任务组合中基线提示词与优化提示词的表现，提升幅度为5.6至25个百分点。上下文方法和对话方法的精确率/召回率达到~90%。C面板展示了优化框架的效率。PO2G只需3次迭代，而基线方法需要6次。D面板从多个维度比较了提示工程与微调。在准确率方面，微调更胜一筹（F1值为97.29%，而提示工程为92.74%）。在部署灵活性和速度方面，提示工程则更具优势。

（Schmidt等人，2025）对 ChatGPT-4o、Claude 和 Gemini 的结构化数据生成进行了比较分析。研究显示，这些模型在准确率、效率和资源使用方面各有显著差异。研究人员发现，Claude 的输出最为准确。Gemini 的各项指标表现均衡，而 ChatGPT-4o 在词元使用量和耗时方面表现最佳。研究结果表明，选择合适的模型取决于现实中的限制和需求，这是因为不同模型在不同维度上的性能特点各不相同。

（Trad 和 Chehab，2024年）探讨了提示工程与微调在网络钓鱼检测中的优缺点，这一点很有参考价值。研究发现，使用提示工程进行测试时，Claude 2 和 GPT-3.5-turbo 的 F1 分数为 92.74%，AUC 值为 97.84%；而经过微调的模型在同一测试集上的得分分别为 97.29% 和 99.56%。性能差距表明，尽管提示工程可以快速部署和使用，但仍可能需要微调。这适用于对精确率要求极高的应用，或高风险用例。在这些情况下，投入额外精力打造更优质的产品是值得的。

实现有效人工智能响应的提示工程类型

提示工程技术可根据其意图和工作原理分为不同类别。从业者了解这些类别后，就能为特定分析任务选择合适的方法，也能组合不同技术，使其协同融入统一的工作流程。

指令式提示这一基础类别，侧重于提供清晰明确的指令，供计算机程序执行特定任务或流程。此类提示词大多采用祈使句式，并通过结构化布局和输出规范来增强清晰度、减少歧义。指令式提示的效果在很大程度上取决于措辞是否精准、指令的组织是否合乎逻辑，以及对输出的限制程度。Hu等（2024年）发现，在实体识别任务中加入系统性指导和错误反馈，可显著提升GPT-3.5在该任务中的临床性能。放宽提示的限制后，MTSamples上的得分从0.634上升至0.794。

上下文提示方法为人工智能提供丰富的背景知识，帮助其理解特定任务或问题的上下文。这种方法对于特定领域应用尤其有用，因为专业知识、技术术语或文化理解都会影响回答。根据(Polak 和 Morgan，2024年)，这项工具展示了上下文提示在 ChatGPT 中的应用潜力。通过大量利用特定领域的上下文以及对话式跟进，他们在材料属性提取方面实现了 90% 的精确率和召回率。回复中的示例、历史背景和技术术语提升了回复的质量和相关性。我们使用“技术”一词，是因为他们在实际应用中使用了这个词。

推理提示通过将推理知识直接嵌入提示词结构，帮助模型提升其复杂推理能力。思维链提示要求明确

识别推理的中间步骤，而树状思维方法则允许同时探索多个解题树。这些技术对于数学计算、逻辑推理及其他多步骤分析任务尤为有效。（Khatuya 等，2025）表明，持续的认知行为指令与灵活的检索机制相结合，相较于基线方法，可将金融问答基准的执行准确率在 FinQA 上提高约 6%，在 ConvFinQA 上提高约 4%。

为确保对话的连贯性，我们采用对话式提示策略。系统会动态管理上下文、保持个性一致性、调整语言模型的沟通风格，并运用其他技术。对话引导的成功与否，取决于管理的精细程度，以及在扩展交互过程中对上下文的保留情况。根据（Polak 和 Morgan，2024），ChatExtract 在访谈提示中系统地提出后续问题，以验证并提高原始材料信息的准确率。

当自然语言处理系统创建新提示词或修改原始提示词以完成任务时，这称为元提示。通过采用递归技术，我们可以提高优化过程的复杂性。此外，通常还能获得比手动处理更稳健的结果。MAPO 会针对每种大语言模型架构微调提示词。MAPO 方法来自（Chen 等，2024）。具体的模型定制使多项任务的准确率稳步提升。

表 2：提示工程类别的特征与效果指标

类别	主要应用	关键特征	报告的有效性范围
指令式	任务规范、格式控制	明确指示，结构化输出	F1值提升：0.16-0.20（临床命名实体识别）
上下文相关	领域特定任务，专业知识	丰富的背景知识，示例、术语	精确率/召回率：~90%（材料提取）
推理	复杂问题求解，数学任务	顺序逻辑，中间步骤	准确率提升：4-6%（金融问答）
对话式	扩展交互，迭代优化	动态上下文，自适应响应	精确率/召回率：~90%（迭代提取）
元提示	自动化优化，模型自适应调优	自我精炼、模型特定适配	持续的任务改进（MAPO 研究）
优化	系统化搜索，效率提升	算法精炼，多目标	减少 3-6 次迭代（PO2G 与基线相比）

根据现有文献，表 2 总结了这些主要提示工程类别的特征、关键用途及报告的有效性范围。

Kusano等，2025 为我们理解提示有效性补充了一个重要细节。不同的提示策略取决于模型及其成本效益，因而各有最佳适用场景。他们针对 23 种提示词、12 个大语言模型以及多个推荐数据集进行了广泛实验。研究发现，高性能模型通常更适合使用简单的提示词，而成本效益更高的模型则能从复杂推理提示和上下文增强中获益。该发现表明，提示工程必须根据模型和任务量身定制。提示工程应考虑模型的能力与局限性。

优化框架与实施策略

系统优化框架能够帮助从业者高效探索复杂且灵活的设计空间，其中可能涉及多个性能标准。用于开发和改进提示工程技术的不同框架包括基于梯度的方法，这类方法用于估计提示有效性的曲面，以及

通过迭代对离散提示词空间进行采样的进化算法，并结合变异与选择。

基于梯度的优化方法将提示优化视为一个连续优化问题，并利用性能梯度近似。据（Lieander 等，2025）的研究，Lieander及其合著者开发的 PO2G 框架效率颇高。经过三轮优化，该框架的准确率达到了约 89%。值得注意的是，用于比较的 ProTeGi 基线也达到了相近的准确率，但需要经过六轮优化。因此，该框架降低了计算成本和算法部署所需的时间，是资源受限应用的可靠选择。

借助批评器和循环对提示的有效性进行微调。Agarwal 等，2024年提出了 PromptWizard，这是一种基于智能体的循环，通过循环中的基于批评器的评估来改进指令和示例。他们在 35 项不同任务上进行了评估，发现其表现始终优于现有提示策略。这表明，智能体驱动方法能够找到有效的提示模式，而这些模式难以通过手动设计或更简单的优化算法发现。

一项重要的改进是模型自适应优化。这体现了这样一种认识：不同的大语言模型架构对各种提示策略的响应并不相同。在 (Chen 等，2024) 中，作者开发了 MAPO，旨在为特定模型设计提示词，而非构造通用提示词。通过使提示词适应各个模型的特性，任务准确率得到了持续提升，这表明针对模型而非任务优化提示词，可能会取得更好的效果。

进化方法和混合方法采用基于种群的方法运用搜索技术。在 2025 年的研究中，(Hazman 等，2025) 将语法引导遗传编程与局部搜索结合，用于离散提示优化。他们在若干小型大语言模型和领域任务上取得了优于相关基线的表现。进化方法的设计工作似乎表明，提示设计空间具有丰富而复杂的结构；与通过梯度引导的局部搜索相比，采用种群进行搜索效果更佳。

多目标优化框架将提示有效性视为可同时从多个维度进行评估的指标。根据 (Narayanaswamy 和 Muniswamy，2025)，NSGA-II 与贝叶斯优化协同作用，以平衡三个目标：准确率、效率和可解释性。基于他们的框架，GLUE 基准实验取得了以下分数：准确率 95%，效率 85%，可解释性 79%。这表明，系统化的多目标优化能够找到在并行评估标准下表现良好的大多数配置。

成本感知优化认识到，通过 API 使用大语言模型需要付费，但其优势必须足以抵消成本。Feng 等，2024 将 GPT-4 所施加的约束下的模式匹配视为 NP-hard 优化问题。他们提出了以一定成本从匹配中尽可能多地获益的方法。Prompt-Matcher 框架在 DeepMDatasets 和 Fabricated-Datasets 上的召回率分别达到 100% 和 91.8%，并明确考虑了计算成本。

现实案例研究与实施成果

各种证据表明，它在多种用途上表现良好。具体而言，它有助于数据科学工作流，也适用于许多现实应用场景。这些案例展示了不同提示策略的有效性，以及这些提示词在该领域有望带来的性能提升。

Polak & Morgan（2024年）开发了 ChatExtract，这是一种基于聊天的系统，可从文献中提取材料属性信息。该方法通过系统化后续询问来检查并完善

提取结果。此类验证和变异流程使精确率和召回率达到 90%以上，使用的模型包括 ChatGPT-4 等顶级提示模型。该系统能够成功从研究论文中提取冷却速率和屈服强度，并以结构化形式将结果存入数据库。若采用传统提取方式，这项工作需要耗费大量人力。这个案例表明，先通过对话式提示进行完善，再进行迭代验证，既能在技术任务中实现高度准确的提取，也能节省大量精力。

根据 Clavié 等人（2023）的研究，可以使用提示工程将招聘信息分为毕业生职位或入门级职位，而提示工程是一种商业智能形式。他们为 GPT-3.5-turbo 创建了一个专门的零样本提示，其95%召回率下的精确率比表现最佳的监督学习基线提高了6%。这意义重大，因为它表明，无需收集训练标签或微调模型，提示工程也能达到监督学习方法的性能。这项研究可能具有一些实际意义，例如缩短实施周期、降低数据收集成本，以及加快分类系统对不断变化的劳动力市场状况的适应速度。

Hu 等人（2024）提出了一种用于临床命名实体识别的新型提示工程框架，正如论文中所述。该框架为这一研究提供了动机，并对各种特定任务的贡献进行了审慎评估。使用完整的提示框架时，模型在 MTSamples 和 VAERS 数据集上的表现会显著提升。GPT-4 的宽松F1分数在 MTSamples 数据集上从 0.804 提升至 0.861，在 VAERS 数据集上则从 0.593 提升至 0.736。在临床环境中，准确识别重要实体可能会影响后续的临床决策和患者安全工具。

Trad 和 Chehab（2024）比较了使用 GPT-3.5-turbo 和 Claude 2 进行网络钓鱼检测时的提示工程与微调。其提示工程方法在包含 1,000 个样本的测试集上取得了 92.74% 的 F1分数。对于零样本方法而言，这一结果相当不错。然而，这些经过微调的专用模型在同一测试集上的 F1 分数为 97.29%，AUC 值为 99.56%，差距明显。该案例研究表明，尽管提示工程能使人工智能模型更快部署，并为应用带来更大灵活性，但追求最高准确率或对安全敏感的应用，仍可能希望通过模型微调来进一步提升性能。

Feng 等人（2024）利用 GPT-4 的能力，设计了分别用于语义匹配和缩写匹配的两种专用提示词，以解决模式匹配问题。通过预算感知选择算法管理 API 成本，其 Prompt-Matcher 框架在 DeepMDatasets 上实现了 100% 的召回率，在 Fabricated-Datasets 上实现了 91.8% 的召回率。该系统将对应关系选择表述为成本约束下的 NP-hard 优化问题，并开发了近似算法，以便在遵守预算限制的同时最大化匹配质量。该案例展示了如何使提示工程与形式化优化方案协同工作，以应对现实世界中的部署约束。

Zhang 等人（2023）评估了在一系列基本数据预处理任务中使用提示工程与大语言模型的效果，这些任务包括错误检测、缺失值填补和实体匹配。Goldwasser等人（2023）在论文《在 WITH数据集上评估 GPT-3.5、GPT-4 和 Vicuna-13B》中，对 GPT-3.5、GPT-4 和 Vicuna-13B 在 12 个不同数据集上的表现进行了比较研究，结果显示，GPT-4 在 4 个数据集上的准确率或 F1分数达到 100%，并且在部分结构化预处理任务中表现良好。不过，不同预处理操作、数据集及其他方面的表现仍存在很大差异。这表明，要有效开展数据预处理，提示工程需要针对具体任务进行精心设计并加以验证。

He等人（2025）研究了在缺少黄金标签的情况下，人类在数据标注任务中的提示工程表现，并发现了实际局限性。他们的研究发现，无指导的人工提示优化可能颇具挑战性，但部分参与者经过数轮迭代后确实有所进步。不过，研究指出，在再经过四轮迭代后，20名参与者中只有9名的标注准确率有所提高

迭代。 这表明，系统化优化框架和自动化提示十分重要；在缺少明确的表现反馈时，仅凭人类直觉不足以识别出优质提示词。

讨论与实践意义

不同场景下的实验表明，提示工程切实有效且可行。提示工程往往取决于具体任务。这意味着，提示工程带来的提升通常幅度较小，但仍然可靠（在某些分类任务中可提升 6%）；而在其他任务中，提升幅度则相当显著（例如，在某些临床命名实体识别应用中可提升 30 个百分点）。换言之，从业者必须针对自己的应用场景实证检验提示工程，而不应假设它在所有场景中都适用。若进一步考虑模型的复杂程度与理想提示策略之间的关系，情况就不那么明朗了。
Kusano 等人（2025年）指出，高质量模型通过简单提示词能取得更好的结果，而低成本模型则能从复杂推理和上下文改进中获益。这意味着，我们应根据模型的具体能力采用提示工程技术，而不是不加区分地应用于不同架构。选择提示工程还是微调来达到预期性能，很大程度上取决于应用的功能、部署所需时间以及性能门槛要求。Trad 和 Chehab（2024年）指出，提示工程能够加快模型部署，并实现可靠的性能（F1 = 92.74%）。然而，微调可能带来显著的额外提升（F1 = 97.29%），这对于关键应用或许不可或缺。企业必须在快速灵活的提示工程与能够带来更佳性能的微调之间作出选择。第四个要素是，系统化优化框架有助于改进手动编写提示。比如，基于梯度的方法和基于智能体的方法能够提升性能，并减少所需的迭代次数。PO2G 经过三次迭代便达到目标性能，而基线方法则需要六次。它还降低了计算成本并缩短了部署时间（Lieander 等人，2025年）。得益于这些效率提升，对持续优化基础设施的投入可为大规模部署提示工程系统的组织带来丰厚回报。在实际应用中，若需要同时从多个维度评估系统效果，那么能够明确展示准确率、效率和/或可解释性之间权衡或平衡关系的多目标优化框架就十分重要。通过多目标优化系统化地结合贝叶斯算法和遗传算法，可以获得提示配置，使准确率达到 95%、效率达到 85%、可解释性达到 79%，从而在复杂的多目标权衡空间中取得颇具竞争力的结果。

局限性与未来研究方向 ns

目前，提示工程的研究与实践面临着各种局限。研究人员使用不同的指标、基准或实验方案。由于缺乏标准化评估方案，这限制了系统性比较。为了获得比较权威并建立在累积知识的基础上，还需要开展更多研究，以制定评估框架。为了获得更好的结果而不时更改文本，是人们面临的一大难题。因此，提示词很难既保持稳定又进行更改。在生产环境中，这种情况更为常见，因为可能更需要更改提示词。微小的扰动需要强有力的提示策略，而研究必须帮助我们制定这些策略。自动化优化框架也难以解释。复杂的优化算法可能会找到一个能产生良好结果的提示词，但我们不知道原因。如果我们能够弄清某些提示词表现良好的原因，就可以学会系统地设计提示词。当实际结果不符合预期时，这也会有所帮助。文化敏感性和偏差缓解应当纳入我们的考虑范围。在一种文化或语言中有用的问题，在另一种文化或语言中可能产生糟糕或带有偏差的回答。探索具有文化适应性的提示策略，并使其在多元群体中保持有效，是重要且符合伦理的研究领域。将提示工程与其他用于增强人工智能的技术（如检索增强生成、工具使用和多智能体系统）结合起来，是一个令人兴奋的研究方向。或许，认识到提示

工程与这些协同技能的关联，可能会带来优于任何单一技术的成果。

结论

提示工程增强了人工智能在广泛的数据科学应用中的功能，同时提升了可衡量的有效性，并降低了采用人工智能的技术障碍。研究表明，提示可将性能提升 6% 至 30% 以上。模型和优化器越复杂，提升效果越显著。与花时间设计人工策划提示相比，采用系统化优化框架更加有效。一种策略通过基于梯度的方法，以更少的步骤达到目标性能；而基于智能体的方法则可能发现性能更好的提示词。实际应用的意义不止于性能指标，还在于增强人工智能的能力、缩短上市时间，并提高适应变化的灵活性。尽管如此，若不谨慎进行任务优化、选择适合模型的策略并开展多参数评估，就无法获得这些益处。随着这一领域的发展，标准化评估框架、稳健的提示策略和可解释优化方法，对于将研究进展成功转化为面向不同利益相关者的可靠有效系统至关重要。

披露声明

作者未报告任何潜在利益冲突

资金支持

T作者报告称，本文所述研究未获得任何资金支持。 le.

作者贡献

所有作者均平等参与了本文的构思、起草、撰写和编辑。每位作者均已审阅并批准最终文本，并同意对本文内容及其准确性负责。

致谢

我谨此感谢 SciSpace 帮助我查找相关文献并进行差距识别，也感谢 Grammarly 帮助提升语言质量。

参考文献

Agarwal, E., Singh, J., Dani, V., Magazine, R., Ganu, T., & Nambi, A.（2024）。*PromptWizard*：任务感知提示优化框架。https://arxiv.org/pdf/2405.18369（预印本，未经同行评审）。

Chen, Y., Wen, Z., Fan, G., Chen, Z., Wu, W., Liu, D., Li, Z., Liu, B., & Xiao, Y.（2024）。MAPO：通过模型自适应提示优化提升大语言模型性能。计算语言学协会 *EMNLP 2023* 论文集，3279–3304。https://doi.org/10.18653/v1/2023.findings-emnlp.215

Clavié, B., Ciceu, A., Naylor, F., Soulié, G., & Brightwell, T. (2023年). 工作场所中的大语言模型：提示工程在职位类型分类中的案例研究。 计算机科学讲义（含人工智能讲义和生物信息学讲义子系列），*13913 LNCS*，3–17。 https://doi.org/10.1007/978-3-031-35320-8_1

Feng, L., Li, H., & Zhang, C. J. (2024年). Prompt-Matcher：利用大模型降低模式匹配结果的不确定性。https://arxiv.org/pdf/2408.14507（预印本，未经同行评审）。Ge, Y., Mei, L., Duan, Z., Li, T., Zheng, Y., Wang, Y., Wang, L., Yao, J., Liu, T., Cai, Y., Bi, B., Guo, F., Guo, J., Liu, S., & Cheng, X. (2025). 大语言模型辅助氛围编程综述。http://arxiv.org/abs/2510.12399（预印本，未经同行评审）。Hazman, M., Pham, M.-K., Soundararajan, S., Mordido, G., Custode, L., Lynch, D., Cruciata, G., Shi, Y., Song, H., Chao, W., Yue, P., Milenovic, A., & Agapitos, A. (2025). 离散提示优化的语法引导进化搜索。https://arxiv.org/pdf/2507.10326v1（预印本，未经同行评审）。He, Z., Naphade, S., & Huang, T. H. K. (2025). 黑暗中的提示：在缺少黄金标签时评估人工进行数据标注提示工程的表现。《计算系统人因会议论文集》。https://doi.org/10.1145/3706598.3714319;CTYPE:STRING:BOOK Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., Zhou, Y., Li, Z., Jiang, X., Lu, Z., Roberts, K., & Xu, H. (2024年). 通过提示工程改进临床命名实体识别的大语言模型。《美国医学信息学协会杂志：JAMIA》，31(9)，1812–1820。https://doi.org/10.1093/JAMIA/OCAD259 Khatuya, S., Naidu, S., Goyal, P., & Ganguly, N. (2025). 用于金融推理的思维程序：利用动态上下文示例和生成式检索。https://arxiv.org/pdf/2510.13157（预印本，未经同行评审）。Kusano, G., Akimoto, K., & Takeoka, K. (2025). 重新审视提示工程：对基于大语言模型的个性化推荐进行全面评估。832–841。https://doi.org/10.1145/3705328.3748159;CSUBTYPE:STRING:CONFERENCE Lieander, A. J., Wang, H., & Rafferty, K. (2025). 面向大语言模型分类任务的双梯度提示优化。《人工智能》2025年，第6卷，第182页，6(8)，182。https://doi.org/10.3390/AI6080182 Narayanaswamy, S. K., & Muniswamy, R. (2025). 使用混合多目标优化框架实现大语言模型的高效提示工程。《控制论与信息技术》，25(2)，67–82。https://doi.org/10.2478/CAIT-2025-0012 Polak, M. P., & Morgan, D. (2024年). 使用对话式语言模型和提示工程从研究论文中提取准确的材料数据。《自然通讯》 2024 年，第 15 卷，第1期，15 1 1 11 10 1038 41467 024 45914 8 : ( ), – 。https://doi.org/ . /s - - - Schmidt, D., Elnashar, A., & Schmidt, D. (2025). 面向结构化数据的提示工程：风格与大语言模型表现的比较评估。https://doi.org/10.20944/PREPRINTS202506.1937.V1（预印本，未经同行评审）。Trad, F., & Chehab, A. (2024年). 提示工程还是微调？使用大语言模型进行网络钓鱼检测的案例研究。《机器学习与知识抽取》2024年，第6卷，第367–384页，6(1)，367–384。https://doi.org/10.3390/MAKE6010018 Zhang, H., Dong, Y., Xiao, C., & Oyamada, M. (2023年). 大语言模型作为数据预处理器。https://arxiv.org/pdf/2308.16361（预印本，未经同行评审）。